Detecting epistatic selection with partially observed genotype data by using copula graphical models

Behrouzi, Pariya; Wit, Ernst C.

Published in:
Journal of the Royal Statistical Society. Series C: Applied Statistics

DOI:
10.1111/rssc.12287

Publication date:
2019

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Applied Statistics
Series C

Appl. Statist. (2019)
68, *Part* 1, *pp.* 141–160

Detecting epistatic selection with partially observed genotype data by using copula graphical models

Pariya Behrouzi

Wageningen University and Research, and University of Groningen, The Netherlands

and Ernst C. Wit

University of Groningen, The Netherlands

[Received November 2016. Revised March 2018]

Summary. In cross-breeding experiments it can be of interest to see whether there are any synergistic effects of certain genes. This could be by being particularly useful or detrimental to the individual. This type of effect involving multiple genes is called epistasis. Epistatic interactions can affect growth, fertility traits or even cause complete lethality. However, detecting epistasis in genomewide studies is challenging as multiple-testing approaches are underpowered. We develop a method for reconstructing an underlying network of genomic signatures of high dimensional epistatic selection from multilocus genotype data. The network captures the conditionally dependent short- and long-range linkage disequilibrium structure and thus reveals 'aberrant' marker–marker associations that are due to epistatic selection rather than gametic linkage. The network estimation relies on penalized Gaussian copula graphical models, which can account for a large number of markers p and a small number of individuals n. We demonstrate the efficiency of the proposed method on simulated data sets as well as on genotyping data in *Arabidopsis thaliana* and maize.

Keywords: Epistasis; Epistatic selection; Gaussian copula; Graphical models; Linkage disequilibrium; Penalized inference

1. Introduction

Recombinant inbred lines (RILs) are a popular study design for studying the genetic and environmental basis of complex traits in biomedical and agricultural research. Many panels of RILs exist in a variety of plant and animal species. RILs are typically derived from two divergent inbred parental strains, but multiparental RILs have been recently applied in *Arabidopsis thaliana, Drosophila* and mouse originating from four or eight inbred parents (Broman, 2005; Gibson and Mackay, 2002; Threadgill *et al*., 2002). The construction of RILs is not always straightforward: low fertility, even complete lethality, of lines during inbreeding is common, particularly in natural outcrossing species (Rongling and Li, 1999; Wu and Li, 2000) and can severely bias genotype frequencies in advanced inbreeding generations. These genomic signatures are indicative of epistatic selection having acted on entire networks of interacting loci during inbreeding, with certain combinations of parental alleles being strongly favoured over others.

Recently, Colomé-Tatché and Johannes (2016) studied two-loci epistatic selection in RILs. The reconstruction of *multiloci* epistatic selection networks, however, has received little attention

Address for correspondence: Ernst C. Wit, Institute of Computational Science, University of Lugano, Via G. Buffi 13, 6900 Lugano, Switzerland.
E-mail: E.C.Wit@rug.nl

by experimentalists. One important reason is that large numbers of potentially interacting loci are methodologically and computationally challenging. One intuitive approach to this problem is performing an exhaustive genome scan for pairs of loci that show significant long-range linkage disequilibrium (LD) or pairwise segregation distortion, and then to try to build up larger networks from overlapping pairs. Törjék *et al*. (2006), for instance, employed this idea for the detection of possible epistasis by testing for pairwise segregation distortion. The drawback of such an approach is that hypothesis testing in the genome scale is heavily underpowered, so weak long-range LD will go undetected, especially after adjusting for multiple testing. Furthermore, pairwise tests are not, statistically speaking, consistent when two conditionally independent loci are mutually dependent on other loci (Whittaker, 2009), and may, therefore, lead to incorrect signatures.

To overcome some of these issues, we shall argue that the detection of epistatic selection in RIL genomes can be achieved by inferring a high dimensional graph of conditional dependence relationships between loci, where the number of markers p can far exceed the number of individuals n. The estimated conditional independence graph captures the conditionally dependent short- and long-range LD structure of RIL genomes and thus provides a basis for identifying associations between distant markers that are due to epistatic selection rather than gametic linkage.

In this paper, we introduce an efficient method to perform this estimation. For this, we propose an l_1-regularized latent graphical model, which involves determining the joint probability distribution of discrete ordinal variables. The genotype data contain information on measured markers in the genome which are generally coded as the number of paternal or maternal alleles, for instance 0, 1 and 2 for a heterozygous population in a diploid species. Sklar's theorem shows that any p-dimensional joint distribution can be decomposed into its p marginal distributions and a copula, which describes the dependence structure between p-dimensional multivariate random variables (Nelsen, 1999). Various statistical network modelling approaches have been proposed for inferring high dimensional associations between non-Gaussian variables (Liu *et al*., 2009, 2012; Dobra and Lenkoski, 2011; Mohammadi *et al*., 2017). The above-mentioned models have some limitations; the first two methods cannot deal with missing data, and the last two are computationally expensive since their inference is based on a Bayesian approach. Studying the conditional relationships between ordinal discrete variables is complicated since we are faced with two challenges. First, general dependence structures can be very complicated, far beyond the pairwise dependences of a normal variate. Second, univariate marginal distributions cannot be adequately described by simple parametric models. To handle the first challenge we use a Gaussian copula, effectively transforming each of the marginal distributions to a standard Gaussian distribution. To address the second challenge, we treat the marginal distributions as nuisance parameters that we estimate non-parametrically.

This paper is organized as follows. In Section 2, we describe the genetic background on epistatic selection. Section 3 explains the model and introduces the Gaussian copula graphical model connecting the observed marker data with the underlying latent genotype. In addition, we explain how to infer the conditional dependence relationships between multiloci in genomewide association studies, using the l_1-regularized Gaussian copula framework. In Section 4, we investigate the performance of the proposed method in terms of precision matrix estimation. Also, we compare the performance of our proposed method with alternative approaches in terms of graph recovery. We have implemented the method in the R package `netgwas` (Behrouzi and Wit, 2017). In Section 5, we aim to reveal genomic regions undergoing selection in two species. We apply our proposed method to the well-studied cross *Col* $\times$ *Cvi* in *Arabidopsis thaliana* in Section 5.1, and to high dimensional *B73* $\times$ *Ki11* genotype data from maize nested association mapping populations in Section 5.2, where 1106 genetic markers were genotyped for 193 individuals.

2. Genetic background of epistatic selection

Two alleles at locations l_1 and l_2 are said to act additively if the effect of the first allele on the phenotype does not depend on the state of the second allele, and vice versa. In contrast, epistasis refers to the interaction of alleles at different loci on that phenotype. Epistasis occurs when the joint effect of a particular pair of loci is different from what would be expected under additivity. In this section, we provide the genetic background on epistatic selection, when the phenotype of interest is survival.

2.1. Meiosis

Sexual reproduction involves meiosis. Meiosis is a form of cell division that produces gametes (egg and sperm). During this process, the arms of homologous chromosomes can recombine, which involves the sequential alignment of genetic material from the maternal and paternal chromosomes. As a result, offspring can have combinations of alleles that are different from those of their parents. Contiguous genetic markers, i.e. particular regions of deoxyribonucleic acid that are close together on the same chromosome, have a tendency to be transmitted together during meiosis. This tendency is called linkage. Loci on different chromosomes have no linkage and they assort independently during meiosis. Statistically speaking, genetic linkage means observing dependence between markers that are physically close on the same chromosome.

LD refers to the coinheritance of alleles at different but functionally related loci. If two loci are in linkage equilibrium, it means that they are inherited completely independently in each generation. If two loci are in LD, it means that certain alleles of each loci are inherited together more or less often than would be expected by chance. This may be due to actual genetic linkage when the loci are on the same chromosome. However, if loci are on different chromosomes, this is due to some form of functional interaction where certain combinations of alleles at two loci affect the viability of potential offspring.

2.2. Recombinant inbred lines

RILs are typically derived by crossing two inbred lines followed by repeated generations of selfing or sibling mating to produce an inbred line whose genome is a mosaic of its parental lines. For instance, if a diploid allele of parent P1 is labelled A and that of P2 is labelled B, then from generation to generation these alleles recombine and produce different genotypes. If due to inbreeding P1 has a homozygous genotype, say A–A (red in Fig. 1), at each locus, whereas P2 has a homozygous genotype, say B–B (green in Fig. 1), at each locus. Crossing P1 and P2 produces an F1 generation with an A–B genotype at each locus. The subsequent F2 followed by repeated selfing results in a genome in the offspring obtained that is a mosaic of the two parental allele combinations that converges to homozygosity at F_∞ (see Fig. 1).

2.3. Genomewide association study

A pure RIL would result in one of two genotype at each locus: either A–A or B–B. However, in practice in a two-way RIL (see Fig. 1), the genotype state of an offspring at a given locus comes either from parent 1, or parent 2 or in a small fraction of cases from both parental alleles. The routine way of coding diploid genotype data is to use $\{0, 1, 2\}$ to represent $\{\mathrm{AA}, \mathrm{AB}, \mathrm{BB}\}$, where we do not distinguish AB and BA.

A complete genome consists of billions of loci, many of which do not vary between individuals in a population. Clearly those loci are inherited without change from generation to generation, unless some mutation occurs. Single-nucleotide polymorphisms (SNPs) are loci

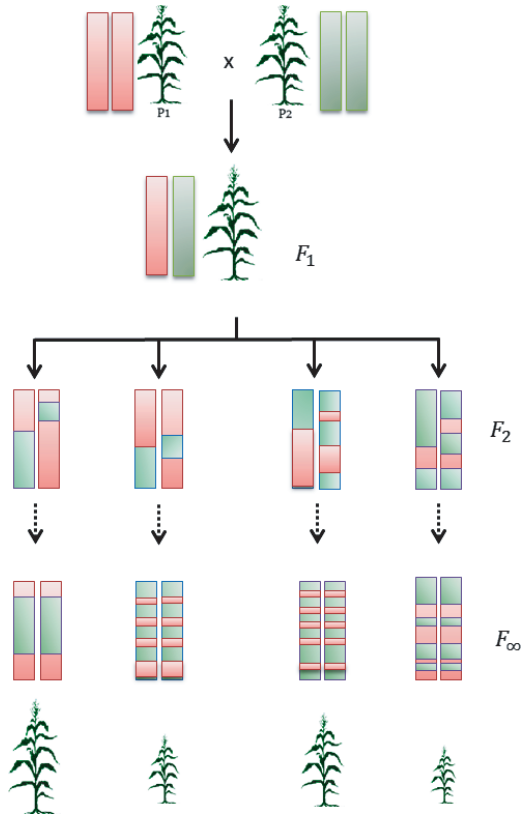


Fig. 1. Production of RILs by repeated selfing

where the genotype does vary, either homozygously $\{0, 2\}$ or heterozygously $\{0, 1, 2\}$, considering diploid organisms. Genomewide association studies measure thousands of SNPs along the genome, resulting for each individual in a vector $Y = (Y_1, \ldots, Y_p)$ of p markers on the genotype. Within each chromosome the markers are ordered, but between chromosomes there is no natural ordering. The component Y_j for an individual indicates the ancestral genotype value for marker j.

Genomewide association studies were designed to identify genetic variations that are associated with a complex trait. In a genomewide association study, typically a small number of individuals are genotyped for hundreds of thousands of SNPs. SNP markers are naturally ordered along the genome with respect to their physical positions. Nearby loci can be highly correlated because of genetic linkage. Moreover, linkage groups typically correspond to chromosomes.

2.4. Epistatic phenotype

Epistasis is typically defined with respect to some explicit phenotype, such as the shape of the comb in a chicken or the flower colour in peas (Bateson, 1909). In RILs the phenotype that we consider, however, is not explicit, but implicit: the viability of the particular genetic recombination of the parental lines results in the presence or absence of such recombination in the progeny.

In the construction of RILs from two divergent parents certain combinations of genotypes may not function well when brought together in the genome of the progeny, thus resulting in sterility, low fertility or even complete lethality of lines during inbreeding. This can result in recombination distortion within chromosomes, short-range LD or segregation distortion across chromosomes, which is also called long-range LD. Thus, the genomic signatures of epistatic selection will appear as interacting loci during inbreeding, whereby some combinations of parental alleles will be strongly favoured over others.

It has long been recognized that detecting the genomic signatures of such high dimensional epistatic selection can be complex, involving multiple loci (Wu and Li, 2000; Mather and Jinks, 1982). The detection of high dimensional epistatic selection is an important goal in population genetics. The aim here is to propose a model for detecting genomic signatures of high dimensional epistatic selection during inbreeding.

3. Graphical model for epistatic selection

If meiosis is a sequential Markov process, then in the absence of epistatic selection the genotype Y can be represented as a graphical model (Lauritzen, 1996) for which the conditional independence graph corresponds to a linear representation of the chromosome structure (Fig. 2(a)). However, in the presence of epistatic selection, the conditional independence of non-neighbouring markers may become undone. This could result, for example, in an underlying conditional independence graph as shown in Fig. 2(b), which shows six markers on two chromosomes whereby markers 2 and 5 have an epistatic interaction that affects the viability of the offspring.

In the next section, we define a convenient semiparametric model, which can easily be generalized to large sets of markers. We let $y_j^{(i)}$, $j=1,\ldots,p$, $i=1,\ldots,n$, denote the genotype of the ith individual for the jth SNP marker. The observations $y_j^{(i)}$ arise from $\{0,1,\ldots,k_j-1\}$, $k_j \geqslant 2$, discrete ordinal values. In the genetic set-up, k_j is the number of possible distinct genotype states at locus j. For instance, in a tetraploid species k_j takes either the value 2 in an inbred homozygous population or 5 in a heterozygous population.

3.1. Gaussian copula graphical model

A copula is a multivariate cumulative distribution function with uniform marginals. Sklar's theorem shows that any p-dimensional joint distribution can be decomposed into its p marginal distributions F_j and a copula C. This decomposition implies that the copula captures the de-

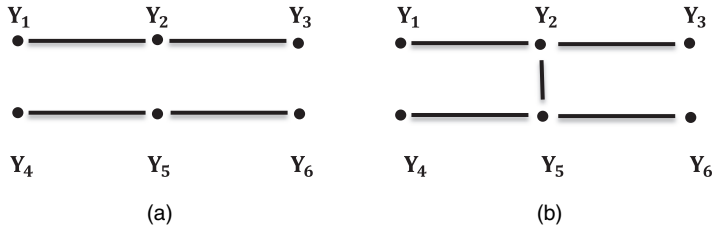


Fig. 2. Schematic representation of six markers on two different chromosomes where Y_1, Y_2 and Y_3 belong to chromosome 1 and Y_4, Y_5 and Y_6 belong to chromosome 2: conditional independence relationships between markers (a) in the absence of epistatic selection, in other words markers on different chromosomes segregate independently, and (b) in the presence of epistatic selection; markers 2 and 5 have an epistatic interaction, resulting in long-range LD

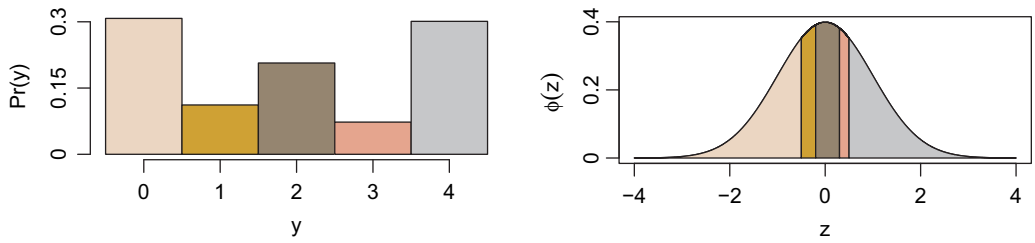


Fig. 3. Relationship between the jth true latent values z_j and the jth observed variable, y_j: here, $k=5$ corresponding to the distinct genotype states in tetraploid species, which contain four copies of the same chromosome

pendence structure between the p variables. Let y be the collection of all p measured genetic markers across a genome. A genetic marker Y_j takes a finite number of ordinal values from $\{0,1,\ldots,k_j-1\}$, with $k_j \geqslant 2$. The marker Y_j is defined as the discretized version of a continuous variable Z_j, which cannot be observed directly. The variable Z helps us to construct the joint distribution of Y as

$$Z \sim N_p(0, \Theta^{-1}),$$

and the Gaussian copula modelling can be expressed as

$$Y_j = F_j^{-1}\{\Phi(Z_j)\},$$

where Θ^{-1} is a correlation matrix for the Gaussian copula, and F_j denotes the univariate distribution of Y_j. We write the joint distribution of Y as

$$P(Y_1 \leqslant y_1, \ldots, Y_p \leqslant y_p) = C\{F_1(y_1), \ldots, F_p(y_p)|\Theta\},$$

where

$$C\{F_1(y_1), \ldots, F_p(y_p)|\Theta^{-1}\} = \Phi_{\Theta^{-1}}[\Phi^{-1}\{F_1(y_1)\}, \ldots, \Phi^{-1}\{F_p(y_p)\}]. \tag{1}$$

Here, Φ defines the cumulative distribution function of the standard univariate normal distribution and Φ_Σ is the cumulative distribution function of $N_p(0,\Sigma)$.

Our aim is to reconstruct the underlying conditional independence graph by using the continuous latent variable Z. Typically the relationship between the jth marker Y_j and the corresponding Z_j is expressed through a set of cut points $-\infty = c_{j,0} < c_{j,1} < \ldots < c_{j,k_j-1} < c_{j,k_j} = \infty$, where $c_{j,y+1} = \Phi^{-1}\{F_j(y)\}$. Thus, $y_j^{(i)}$ can be written as

$$y_j^{(i)} = \sum_{l=0}^{k_j-1} l \times I_{\{c_{j,l} < z_j^{(i)} \leqslant c_{j,l+1}\}}, \qquad i = 1,2,\ldots,n. \tag{2}$$

The jth observed variable $y_j^{(i)}$ takes its value according to latent variable $z_j^{(i)}$. Fig. 3 displays how the observed data can be obtained from the latent variable by using the Gaussian copula.

Assuming that $\mathcal{D}_F(y) = \{z_j^{(i)} \in \mathbb{R} | c_{j,y_j^{(i)}} < z_j^{(i)} \leqslant c_{j,y_j^{(i)}+1}\}$, the likelihood function of a given graph with a precision matrix Θ and marginal distributions F is defined as

$$L_y(\Theta, F) = \int_{\mathcal{D}_F(y)} p(z|\boldsymbol{\Theta}) \mathrm{d}z. \tag{3}$$

3.2. l_1-penalized inference of Gaussian copula graphical model

Let $y^{(1)}, \ldots, y^{(n)}$ be independent and identically distributed sample values from the above Gaussian copula distribution. A copula formulation enables us to learn the marginals F_j separately from the dependence structure of p-variate random variables. Our aim is to estimate the precision matrix and we treat the marginals as nuisance parameters and estimate them non-parametrically through the empirical distribution function $\hat{F}_j(y) = (1/n)\Sigma_{i=1}^{n} I\{y_j^{(i)} \leqslant y\}$. Hence, in likelihood (3) the precision matrix of the Gaussian copula, Θ, is the only parameter to estimate, as we replace $\mathcal{D}_F(y)$ by $\mathcal{D}_{\hat{F}}(y)$ which we shall simply indicate as $\hat{\mathcal{D}}$.

We impose a sparsity penalty on the elements of the precision matrix $\boldsymbol{\Theta}$ by using an l_1-norm penalty (Abegaz and Wit, 2015; Friedman *et al.*, 2008). Genetically speaking, this sparsity is sensible as we expect *a priori* only a small number of pairs of LD markers beyond the neighbouring markers. The l_1-penalized log-likelihood function of genetic markers can be written as

$$l_{\mathbf{y}}^{p}(\boldsymbol{\Theta}) \approx \frac{n}{2}\log|\boldsymbol{\Theta}| - \frac{1}{2}\sum_{i=1}^{n}\int_{\hat{\mathcal{D}}}\cdots\int z^{(i)^{\mathrm{T}}}\boldsymbol{\Theta}z^{(i)}\mathrm{d}z_1^{(i)}\ldots\mathrm{d}z_p^{(i)} - \lambda\|\Theta\|_1, \tag{4}$$

where $z^{(i)} = (z_1^{(i)}, \ldots, z_p^{(i)})^{\mathrm{T}}$. The maximum $\hat{\Theta}_\lambda$ of this log-likelihood function has no closed form expression. To address this problem we introduce a penalized expectation–maximization (EM) algorithm.

The penalized EM algorithm proceeds by iteratively computing in the E-step the conditional expectation of the joint log-likelihood and optimizing this conditional expectation in the M-step. Assuming that $\hat{\boldsymbol{\Theta}}_\lambda^{(m)}$ is the updated approximation of $\hat{\boldsymbol{\Theta}}_\lambda$ in the M-step, then in the E-step the conditional expectation of the joint penalized log-likelihood given the data and $\hat{\boldsymbol{\Theta}}^{(m)}$ is determined:

$$\begin{aligned} Q(\boldsymbol{\Theta}|\hat{\boldsymbol{\Theta}}^{(m)}) &= E_Z\left[\sum_{i=1}^{n}\log\{p(Z^{(i)}|\boldsymbol{\Theta})\}|y^{(i)}, \hat{\boldsymbol{\Theta}}^{(m)}, \hat{\mathcal{D}}\right] \\ &= \frac{n}{2}\left[\log|\boldsymbol{\Theta}| - \mathrm{tr}\left\{\frac{1}{n}\sum_{i=1}^{n}E_{Z^{(i)}}(Z^{(i)}Z^{(i)^{\mathrm{T}}}|y^{(i)}, \hat{\boldsymbol{\Theta}}^{(m)}, \hat{\mathcal{D}})\boldsymbol{\Theta}\right\} - p\log(2\pi)\right], \end{aligned} \tag{5}$$

and

$$Q_\lambda(\boldsymbol{\Theta}|\hat{\boldsymbol{\Theta}}^{(m)}) = Q(\boldsymbol{\Theta}|\hat{\boldsymbol{\Theta}}^{(m)}) - \lambda\|\Theta\|_1.$$

In this equation we still need to evaluate $\bar{R} = (1/n)\Sigma_{i=1}^{n}E(Z^{(i)}Z^{(i)^{\mathrm{T}}}|y^{(i)}, \hat{\boldsymbol{\Theta}}^{(m)}, \hat{\mathcal{D}})$, which we do via one of the two following approaches.

3.2.1. Monte Carlo Gibbs sampling of latent covariance

In the Gibbs sampling technique we randomly generate for each sample $Y^{(i)}$ a number of Gibbs samples $Z_*^{(i)1}, \ldots, Z_*^{(i)N}$ from a p-variate truncated normal distribution, whose boundaries come from the cut points of $Y^{(i)}$, as implemented in the R package `tmvnorm` (Geweke, 2005). Let

$$Z_*^{(i)} = \begin{pmatrix} Z_*^{(i)1} \\ \vdots \\ Z_*^{(i)N} \end{pmatrix} \in \mathbb{R}^{N\times p}$$

represent the Gibbs samples for each sample in the data. The expected individual covariance matrix $R_i = E(Z^{(i)}Z^{(i)^{\mathrm{T}}}|y^{(i)}, \hat{\Theta}^{(m)}, \hat{\mathcal{D}})$ can then be estimated as

$$\hat{R}_i = \frac{1}{N} Z_*^{(i)^{\mathrm{T}}} Z_*^{(i)}.$$

To estimate $\bar{R}$ we take the average of the individual expectation $\hat{\bar{R}} = (1/n)\Sigma_{i=1}^{n} \hat{R}_i$. We remark that $\hat{\bar{R}}$ is a positive definite matrix with probability 1 as long as $N \geq p/n$. As $Z_{*l}^{(i)^{\mathrm{T}}} Z_{*l}^{(i)}$ is a rank 1 non-negative definite matrix with probability 1 and, therefore, $\hat{\bar{R}}$ is of full rank and strictly positive definite with probability 1, where $Z_{*l}^{(i)}$ is the lth row of $Z_*^{(i)}$. In practice, we noted that the Gibbs sampler needs only a few burn-in samples and $N = 1000$ sweeps are sufficient to calculate the mean of the conditional expectation accurately (more details are given in the on-line supplementary materials).

3.2.2. *Approximation of the conditional expectation*

Alternatively, we use an efficient approximate estimation algorithm (Guo *et al.*, 2015). The variance elements in the conditional expectation matrix can be calculated through the second moment of the conditional $z_j^{(i)}|y^{(i)}$, and the rest of the elements in this matrix can be approximated through $E(Z_j^{(i)} Z_{j'}^{(i)}|y^{(i)}; \hat{\Theta}, \hat{\mathcal{D}}) \approx E(Z_j^{(i)}|y^{(i)}; \hat{\Theta}, \hat{\mathcal{D}}) E(Z_{j'}^{(i)}|y^{(i)}; \hat{\Theta}, \hat{\mathcal{D}})$ by using mean field theory (Peterson, 1987). The first and second moments of $z_j^{(i)}|y^{(i)}$ can be written as

$$E(Z_j^{(i)}|y^{(i)}, \hat{\Theta}, \hat{\mathcal{D}}) = E\{E(Z_j^{(i)}|z_{-j}^{(i)}, y_j^{(i)}, \hat{\Theta}, \hat{\mathcal{D}})|y^{(i)}, \hat{\Theta}, \hat{\mathcal{D}}\}, \tag{6}$$

$$E\{(Z_j^{(i)})^2|y^{(i)}, \hat{\Theta}, \hat{\mathcal{D}}\} = E[E\{(Z_j^{(i)})^2|z_{-j}^{(i)}, y_j^{(i)}, \hat{\Theta}, \hat{\mathcal{D}}\}|y^{(i)}, \hat{\Theta}, \hat{\mathcal{D}}], \tag{7}$$

where $z_{-j}^{(i)} = (z_1^{(i)}, \ldots, z_{j-1}^{(i)}, z_{j+1}^{(i)}, \ldots, z_p^{(i)})$. The inner expectations in equations (6) and (7) are relatively straightforward to calculate. $z_j^{(i)}|z_{-j}^{(i)}, y_j^{(i)}$ follows a truncated Gaussian distribution on the interval $[c_{y_j^{(i)}}^{(j)}, c_{y_j^{(i)}+1}^{(j)}]$ with parameters $\mu_{i,j}$ and $\sigma_{i,j}^2$ given by

$$\mu_{ij} = \hat{\boldsymbol{\Sigma}}_{j,-j} \hat{\boldsymbol{\Sigma}}_{-j,-j}^{-1} z_{-j}^{(i)^{\mathrm{T}}},$$

$$\sigma_{i,j}^2 = 1 - \hat{\boldsymbol{\Sigma}}_{j,-j} \hat{\boldsymbol{\Sigma}}_{-j,-j}^{-1} \hat{\boldsymbol{\Sigma}}_{-j,-j}.$$

Let $r_{k,l} = (1/n)\Sigma_{i=1}^{n} E(Z_k^{(i)} Z_l^{(i)}|y^{(i)}, \hat{\boldsymbol{\Theta}}, \hat{\mathcal{D}})$ be the (k,l)th element of empirical correlation matrix $\bar{R}$. Then to obtain $\bar{R}$ two simplifications are required:

$$\begin{aligned} E(Z_k^{(i)} Z_l^{(i)^{\mathrm{T}}}|y^{(i)}, \hat{\boldsymbol{\Theta}}, \hat{\mathcal{D}}) &\approx E(Z_k^{(i)}|y^{(i)}, \hat{\boldsymbol{\Theta}}, \hat{\mathcal{D}}) E(Z_l^{(i)}|y^{(i)}, \hat{\boldsymbol{\Theta}}, \hat{\mathcal{D}}) && \text{if } 1 \leqslant k \neq l \leqslant p, \\ E(Z_k^{(i)} Z_l^{(i)^{\mathrm{T}}}|y^{(i)}, \hat{\boldsymbol{\Theta}}, \hat{\mathcal{D}}) &= E\{(Z_k^{(i)})^2|y^{(i)}, \hat{\boldsymbol{\Theta}}, \hat{\mathcal{D}}\} && \text{if } k = l. \end{aligned}$$

Applying the results in Appendix A to the conditional $z_j^{(i)}|z_{-j}^{(i)}, y_j^{(i)}$ we obtain

$$E(Z_j^{(i)}|y^{(i)}; \hat{\boldsymbol{\Theta}}, \hat{\mathcal{D}}) = \hat{\boldsymbol{\Sigma}}_{j,-j} \hat{\boldsymbol{\Sigma}}_{-j,-j}^{-1} E(Z_{-j}^{(i)^{\mathrm{T}}}|y^{(i)}; \hat{\boldsymbol{\Theta}}, \hat{\mathcal{D}}) + \frac{\phi(\hat{\delta}_{j,y_j^{(i)}}^{(i)}) - \phi(\tilde{\delta}_{j,y_j^{(i)}+1}^{(i)})}{\Phi(\tilde{\delta}_{j,y_j^{(i)}+1}^{(i)}) - \Phi(\tilde{\delta}_{j,y_j^{(i)}}^{(i)})} \tilde{\sigma}_j^{(i)}, \tag{8}$$

$$\begin{aligned} E\{(Z_j^{(i)})^2|y^{(i)}; \hat{\boldsymbol{\Theta}}, \hat{\mathcal{D}}\} =& \hat{\boldsymbol{\Sigma}}_{j,-j} \hat{\boldsymbol{\Sigma}}_{-j,-j}^{-1} E(Z_{-j}^{(i)^{\mathrm{T}}} Z_{-j}^{(i)}|y^{(i)}; \hat{\boldsymbol{\Theta}}, \hat{\mathcal{D}}) \hat{\boldsymbol{\Sigma}}_{-j,-j}^{-1} \hat{\boldsymbol{\Sigma}}_{j,-j}^{\mathrm{T}} + (\tilde{\sigma}_j^{(i)})^2 \\ &+ 2 \frac{\phi(\tilde{\delta}_{j,y_j^{(i)}}^{(i)}) - \phi(\tilde{\delta}_{j,y_j^{(i)}+1}^{(i)})}{\Phi(\tilde{\delta}_{j,y_j^{(i)}+1}^{(i)}) - \Phi(\tilde{\delta}_{j,y_j^{(i)}}^{(i)})} \{\hat{\boldsymbol{\Sigma}}_{j,-j} \hat{\boldsymbol{\Sigma}}_{-j,-j}^{-1} E(Z_{-j}^{(i)^{\mathrm{T}}}|y^{(i)}; \hat{\boldsymbol{\Theta}}, \hat{\mathcal{D}})\} \tilde{\sigma}_j^{(i)} \\ &+ \frac{\delta_{j,y_j^{(i)}}^{(i)} \phi(\tilde{\delta}_{j,y_j^{(i)}}^{(i)}) - \tilde{\delta}_{j,y_j^{(i)}+1}^{(i)} \phi(\tilde{\delta}_{j,y_j^{(i)}+1}^{(i)})}{\Phi(\tilde{\delta}_{j,y_j^{(i)}+1}^{(i)}) - \Phi(\tilde{\delta}_{j,y_j^{(i)}}^{(i)})} (\tilde{\sigma}_j^{(i)})^2, \end{aligned} \tag{9}$$

where $Z_{-j}^{(i)} = (Z_1^{(i)}, \ldots, Z_{j-1}^{(i)}, Z_{j+1}^{(i)}, \ldots, Z_p^{(i)})$ and $\tilde{\delta}_{j,y_j^{(i)}}^{(i)} = \{c_j^{(i)} - E(\tilde{\mu}_{ij}|y^{(i)}; \hat{\boldsymbol{\Theta}}, \hat{\mathcal{D}})\}/\tilde{\sigma}_{ij}$.
In this way, an approximation for $\bar{R}$ is obtained as follows:

$$\tilde{r}_{kl} = \begin{cases} \dfrac{1}{n}\displaystyle\sum_{i=1}^{i=n} E(Z_k^{(i)}|y^{(i)}, \hat{\boldsymbol{\Theta}}^{(m)}, \hat{\mathcal{D}}) E(Z_l^{(i)}|y^{(i)}, \hat{\boldsymbol{\Theta}}^{(m)}, \hat{\mathcal{D}}) & \text{if } 1 \leqslant k \neq l \leqslant p, \\ \dfrac{1}{n}\displaystyle\sum_{i=1}^{i=n} E\{(Z_k^{(i)})^2|y^{(i)}, \hat{\boldsymbol{\Theta}}^{(m)}, \hat{\mathcal{D}}\} & \text{if } k = l. \end{cases}$$

3.2.3. M-step

The M-step involves updating $\boldsymbol{\Theta}$ by maximizing the expected complete likelihood with an l_1-penalty over the precision matrix:

$$\hat{\boldsymbol{\Theta}}_\lambda^{(m+1)} = \arg\max_{\boldsymbol{\Theta}} \{\log|\boldsymbol{\Theta}| - \mathrm{tr}(\bar{R}\boldsymbol{\Theta}) - \lambda\|\boldsymbol{\Theta}\|_1\}.$$

In our implementation, we use the `glasso` method for optimization (Witten *et al.*, 2011). A multicore implementation of our proposed methods speeds up the computational challenge, as all the penalized optimizations are performed in parallel across the available nodes in any multicore computer architecture. This feature proportionally reduces the computational time. Performing simulations, we noted that the EM algorithm converges quickly, i.e. within at most 10 iterations.

3.3. Selection of the tuning parameter

The penalized log-likelihood method guarantees with probability 1 that the precision matrix is positive definite. In addition, the method leads to a sparse estimator of the precision matrix, which encodes the latent conditional independences between the genetic markers. A grid of regularization parameters $\Lambda = (\lambda_1, \ldots, \lambda_N)$ determines the level of sparsity of the precision matrix. Since we are interested in graph estimation, one approach is to subsample the data, to measure the instability of the edges across the subsamples and to choose a λ whose instability is less than a certain cut point value, which is usually taken as 0.05 (Liu *et al.*, 2010). However, in high dimensional settings this approach is time consuming.

Alternatively, we compute various information criteria at EM convergence based on the observed log-likelihood, which can be written as (Ibrahim *et al.*, 2008)

$$l_y(\hat{\Theta}_\lambda) = Q(\hat{\Theta}_\lambda|\hat{\Theta}^{(m)}) - H(\hat{\Theta}_\lambda|\hat{\Theta}^{(m)}), \tag{10}$$

where Q is defined in equation (5) and the H-function is

$$H(\hat{\Theta}_\lambda|\hat{\Theta}_\lambda^{(m)}) = E_z\{l_{Z|Y}(\hat{\Theta}_\lambda)|Y; \hat{\Theta}_\lambda\} = E_z[\log\{f(z)\}|Y; \hat{\Theta}_\lambda] - \log\{p(y)\}.$$

We consider the class of model selection criteria given by

$$\mathrm{IC}_{H,Q}(\lambda) = -2l_{z\in\mathcal{D}}(\hat{\Theta}_\lambda) + \mathrm{bias}(\hat{\Theta}_\lambda).$$

Different forms of the $\mathrm{bias}(\hat{\Theta}_\lambda)$ lead to different information criteria for model selection (Vujačić *et al.*, 2015). As we are interested in graph estimation, we use the extended Bayesian information criterion eBIC that has been introduced for conditional independence graph selection (Foygel and Drton, 2010):

$$\mathrm{eBIC}(\lambda) = -2l(\hat{\Theta}_\lambda) + \{\log(n) + 4\gamma\log(p)\}\,\mathrm{df}(\lambda),$$

where $\mathrm{df}(\lambda) = \Sigma_{1 \leqslant k < l \leqslant p} I(\hat{\Theta}_\lambda \neq 0)$ refers to the number of non-zero off-diagonal elements of $\hat{\Theta}_\lambda$ and $\gamma \in [0, 1]$ is the parameter that penalizes the number of models, which increases when p increases. In the case of $\gamma = 0$ the classical Bayesian information criterian BIC is obtained. Typical values for γ are $\frac{1}{2}$ and 1. To obtain the optimal model in terms of graph estimation we pick the penalty term that minimizes eBIC over $\lambda > 0$.

3.4. Inference uncertainty

The classical likelihood-based method to estimate uncertainty by inverting the Fisher information matrix does not directly apply to penalized likelihood approaches (Lehmann and Casella, 2006). Instead, one way to compute the uncertainty that is associated with the estimation of the precision matrix under the penalized Gaussian copula graphical model is through a non-parametric bootstrap. For the penalized likelihood bootstrap, we replicate B data sets that are created by sampling with replacement n samples from the data set $Y_{n \times p}$. We treat each replicate as the original data and run the entire inference procedure of the proposed Gaussian copula graphical model to estimate $\tilde{\Theta}_\lambda^{(b)}$ $(b = 1, \ldots, B)$. In this bootstrap, we take into account the uncertainty arising from both the empirical estimation of the marginals and the selection of the tuning parameter. Thus, the above-mentioned non-parametric bootstrap procedure adequately reflects the underlying uncertainty in the estimation procedure of the proposed epistatic interaction graph. We have implemented this procedure to evaluate the uncertainty that is associated with the estimation of the epistatic interactions in the *Arabidopsis thaliana* experiment in Section 5.1.2.

4. Simulation study

We study the performance of the proposed method in a simulation study, mimicking the small genotyping study involving *Arabidopsis thaliana* and the medium-sized study involving maize. For each dimension, we consider two different scenarios: in one scenario the latent variables satisfy the multivariate Gaussian distribution, and in the other scenario they do not. In the latter, we consider the t-distribution with 3 degrees of freedom. The simulated data are obtained by different scenarios for $p = 90, 1000$ variables, $n = 200, 360$ samples and $k = 3$ genotype states.

The simulated graphs are selected in such a way as to mimic a realistic epistatic selection network. First, we partition the variables into g linkage groups (each of which represents a chromosome); then within each linkage group adjacent markers are linked via an edge because of genetic linkage. Also, with probability 0.01 a pair of non-adjacent markers in the same chromosome is given an edge. Trans-chromosomal edges are simulated with probability 0.02. In the low dimension case ($p = 90$) we create five chromosomes and in the high dimension case ($p = 1000$) 10 chromosomes. The corresponding positive definite precision matrix Θ has a 0 pattern corresponding to the non-present edges. For each simulation a new random-precision matrix is generated. The latent variables are simulated from either a multivariate normal distribution, $N_p(0, \Theta^{-1})$, or a multivariate t-distribution with 3 degrees of freedom and covariance matrix Θ^{-1}. We generate random cut-off points from a uniform distribution, discretizing the latent space into $k = 3$ disjoint states.

We compare our proposed method with other approaches in terms of receiver operating characteristic performance. Also, we compare our model with other methods in terms of graph recovery. We compare the following four methods:

(a) the proposed method using the Gibbs sampler within the EM algorithm (Gibbs);
(b) the proposed method using numeric approximation within the EM algorithm (Approx);

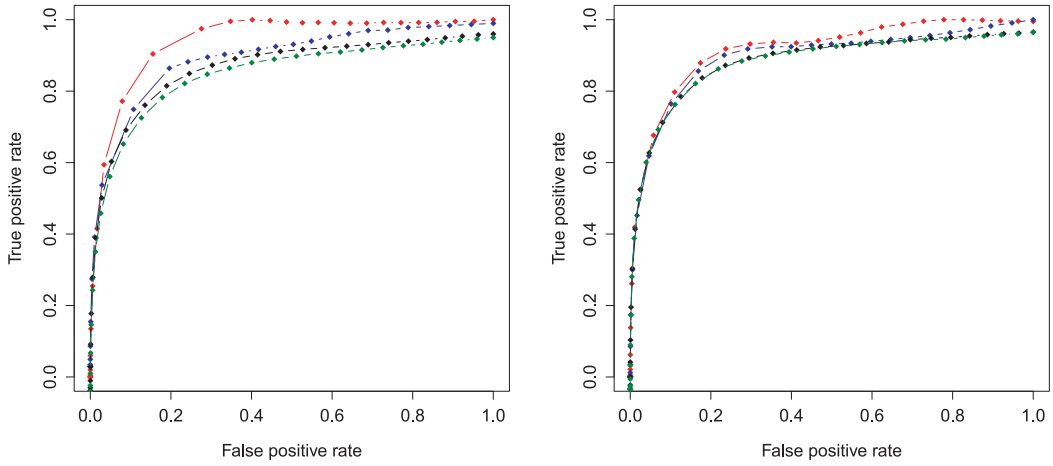


Fig. 4. Receiver operating characteristic curves for comparing different methods of recovering the true graph where $p = 1000$, $n = 100$ and $k = 3$ (——, Gibbs; ——, Approx; ——, NPN-tau; ——, NPN-ns): the data are simulated from (a) the Gaussian copula graphical model and (b) the $t_{(3)}$-copula graphical model; our method (Gibbs) consistently outperforms the other methods

(c) a non-paranormal sceptic using Kendall's τ (NPN-tau) (Liu *et al.*, 2012);
(d) a non-paranormal normal score (NPN-ns) (Liu *et al.*, 2009).

The receiver operating characteristic curves in Fig. 4 show the performance of the various graph estimation methods over 75 repeated simulations each at 30 grid points of the tuning parameter. The area under the curve is used as a measure of the quality of the methods in recovering the true graph. In Fig. 4(a) the latent variable is normally distributed, whereas in Fig. 4(b) the latent variable has a t_3-distribution. Fig. 4 shows how our proposed method, particularly that employing the Gibbs sampler, outperforms the non-paranormal approaches. This is true both in the scenario of no model misspecification, i.e. when the data are simulated from the Gaussian copula graphical model, as well as in the case of model misspecification, i.e. when the data are simulated from a Student $t_{(3)}$-copula graphical model. Our method combined with the approximation approach performs somewhat better than both non-paranormal approaches under both scenarios. Based on our simulation studies the performances of methods NPN-tau and NPN-ns are similar in the absence of outliers, as discussed in Liu *et al.* (2012).

Furthermore, we measure the methods' performance in terms of graph accuracy and its closeness to the true graph. The above penalized inference methods are a path estimation procedure; however, in practice, the practitioner wants to select a particular network. As we are interested in the global true interaction structure, but not in the individual parameters, the extended Bayesian information criterion eBIC is particularly appropriate. To evaluate the accuracy of the estimated graph we compute the F_1-score $2\mathrm{TP}/(2\mathrm{TP} + \mathrm{FP} + \mathrm{FN})$, sensitivity $\mathrm{SEN} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN})$ and specificity $\mathrm{SPE} = \mathrm{TN}/(\mathrm{TN} + \mathrm{FP})$, where TP, TN, FP and FN are the numbers of true positive, true negative, false positive and false negative values respectively, in identifying the non-zero elements in the precision matrix. We note that high values of the F_1-score, sensitivity and specificity indicate good performance of the proposed approach for the given combination of p, n and k. However, as there is a natural trade-off between sensitivity and specificity, we focus particularly on the F_1-score to evaluate the performance of each method. For each simulated data set, we apply each of the four methods.

Table 1. Comparison between the performance of the proposed regularized approximate EM, regularized Gibbs sampler EM, the non-paranormal sceptic Kendall's τ and the non-paranormal normal score†

Statistic	*Results for p = 90, n = 360, k = 3*		*Results for p = 1000, n = 200, k = 3*	
	Normal	$t_{(3)}$	*Normal*	$t_{(3)}$
Gibbs				
F_1-oracle	0.83 (0.02)	0.83 (0.02)	0.75 (0.04)	0.76 (0.02)
F_1	*0.76* (0.03)	*0.75* (0.03)	*0.74* (0.04)	*0.50* (0.06)
SEN	0.97 (0.02)	0.98 (0.01)	0.67 (0.07)	0.26 (0.05)
SPE	0.97 (0.00)	0.97 (0.00)	0.99 (0.00)	0.99 (0.00)
Approx				
F_1-oracle	0.80 (0.02)	0.80 (0.02)	0.73 (0.03)	0.74 (0.02)
F_1	0.70 (0.03)	0.70 (0.03)	*0.73* (0.03)	*0.50* (0.35)
SEN	0.98 (0.02)	0.96 (0.01)	0.70 (0.08)	0.50 (0.35)
SPE	0.96 (0.01)	0.98 (0.01)	1.00 (0.00)	1.00 (0.00)
NPN-tau				
F_1-oracle	0.84 (0.02)	0.84 (0.02)	0.76 (0.03)	0.76 (0.02)
F_1	0.70 (0.15)	0.70 (0.15)	0.00 (0.00)	0.00 (0.00)
SEN	0.94 (0.19)	0.94 (0.19)	0.00 (0.00)	0.00 (0.00)
SPE	0.97 (0.01)	0.97 (0.01)	1.00 (0.00)	1.00 (0.00)
NPN-ns				
F_1-oracle	0.83 (0.02)	0.83 (0.02)	0.75 (0.03)	0.75 (0.03)
F_1	0.65 (0.25)	0.56 (0.32)	0.00 (0.00)	0.00 (0.00)
SEN	0.86 (0.32)	0.74 (0.42)	0.00 (0.00)	0.00 (0.00)
SPE	0.97 (0.01)	0.98 (0.01)	1.00 (0.00)	1.00 (0.00)

†The means of the F_1-score, sensitivity and specificity over 75 replications are represented as selected by using eBIC. Higher values of the F_1-score indicate better network recovery, balancing both sensitivity and specificity. The best method in each column is indicated by italics.

In Table 1, we compare these four methods in a low dimensional case $p=90$, $n=360$ and $k=3$, mimicking the *Arabidopsis thaliana* data set that we consider later, and a high dimensional case of $p=1000$, $n=200$ and $k=3$, mimicking the maize genotype data. In both cases we consider two different scenarios: in one scenario the latent variable satisfies a Gaussian distribution and in the other scenario it is overdispersed according to a $t_{(3)}$-distribution. We report the average values for the F_1-score, SEN and SPE in 75 independent simulations. The value of the F_1-oracle indicates the best values of the F_1-score that can be obtained by selecting the optimal tuning parameter in the l_1-optimization. Table 1 shows that the method proposed using either Gibbs sampling or the approximation method within the EM algorithm performs very well in selecting the best graph. In both scenarios in the low dimension case, method NPN-tau chooses a better graph compared with NPN-ns. However, in the high dimensional case neither of them chooses anything close to the true graph. In fact, they select an empty graph. In contrast, the proposed method, though selecting a relatively sparse graph, finds a considerable fraction of the true links.

We performed all computations on a cluster with 24 Intel Xeon 2.5-GHz cores processor and 128 Gbytes random-access memory. In our method it is possible to estimate the conditional expectation either through Gibbs sampling or using the approximation approach. For large numbers of markers ($p \geqslant 2000$) the Gibbs sampling approach is not recommended because of excessive computational costs. However, the approximation approach can handle such situa-

Table 2. Computational cost comparison between the method proposed (Approx) and the non-paranormal sceptic method NPN-tau†

Method	*Computational cost (min) for the following numbers of variables:*					
	100	*500*	*1000*	*2000*	*3000*	*4000*
Approx	0.34	1.26	19.71	80.43	734.79	2623.68
NPN-tau	0.03	0.16	1.76	14.05	62.76	—‡

†For the larger *p*s the non-paranormal sceptic method is faster than our proposed method. However, neither NPN-tau nor NPN-ns can deal with missing values, whereas the approximation approach is developed to be able to deal with missing genotypes that commonly occur in genotype data.
‡Exceeds the step memory limit at some point.

tions. The computational costs for the non-paranormal sceptic and the normal score methods are similar to each other. Therefore, in Table 2 we report the computational cost of the approximation method proposed and the non-paranormal sceptic method *versus* the number of variables for a sample size fixed at 200. Both methods have roughly similar increases in computational time, which seems to be larger than quadratic in p. Our method is roughly a constant factor of 10 times more than the times for the non-paranormal sceptic. This is due to the EM iterations. The EM algorithm has advantages, however, as our method can deal with missing genotype values, which are very common in practice. Moreover, by implementing the algorithm in multicore we have significantly reduced the computational costs. Further improvement can be achieved by reprogramming the algorithm in C++ and interfacing it with R.

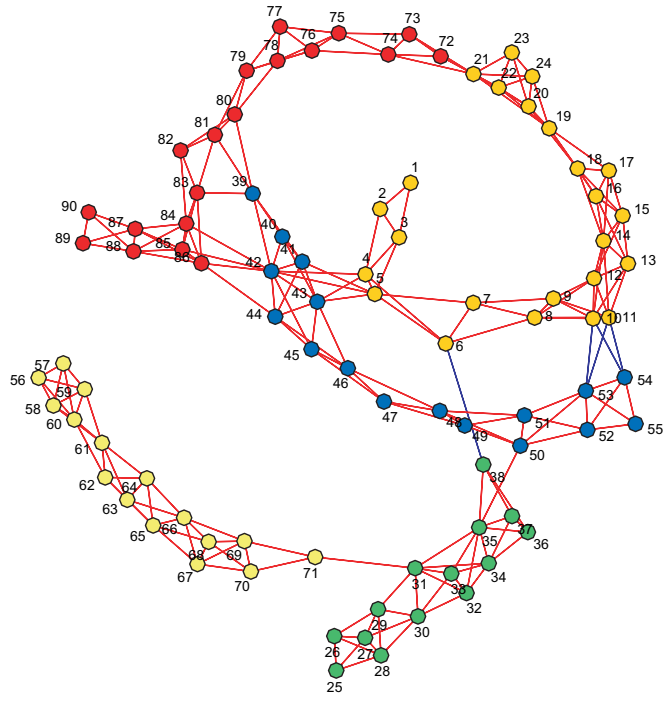
5. Detecting genomic signatures of epistatic selection

5.1. Epistatic selection in Arabidopsis thaliana

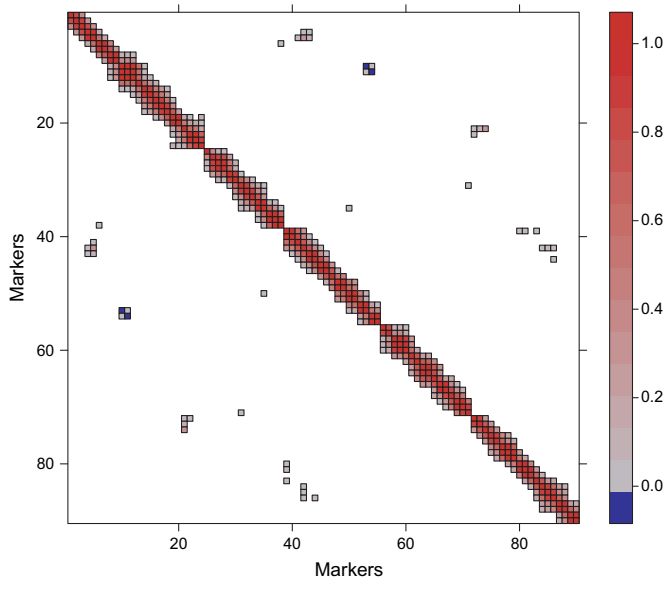
We apply our proposed Gibbs sampling approach to detect epistatic selection in *Arabidopsis thaliana* genotype data that are derived from an RIL cross between Columbia-0 (*Col-0*) and the Cape Verde Island (*Cvi-0*), in which 367 individual plants were genotyped across 90 genetic markers (Simon *et al.*, 2008). The *Cvi-0* × *Col-0* RIL is a diploid population with $k = 3$ possible genotypes. The genotype data are coded as $\{0, 1, 2\}$, where 0 and 2 represent homozygous genotypes from *Col-0* and *Cvi-0* respectively, whereas 1 defines the heterozygous genotype. Some markers have missing genotypes (0.2%).

The results of the analysis are presented in Fig. 5. If there is no LD, markers in different chromosomes should segregate independently. The first thing to note is that the Gaussian copula graphical model automatically groups markers that belong to the same chromosome, because of genetic linkage. In the diagonal of Fig. 5(b) the five chromosomes of *Arabidopsis thaliana* plant are clearly identifiable.

Existence of trans-chromosomal conditional dependences reveals the genomic signatures of epistatic selection. Fig. 5 shows that our method finds some trans-chromosomal regions that do interact. In particular, the bottom of chromosome 1 and the top of chromosome 5 do not segregate independently of each other. Beside this, interactions between the tops of chromosomes 1 and 3 involve pairs of loci that also do not segregate independently. This genotype has been studied extensively in Bikard *et al.* (2009). They reported that the former interaction that causes



(a)



(b)

Fig. 5. Inferred network for genotype data in a cross between the *Arabidopsis thaliana* accessions, Columbia-0 (*Col-0*) and Cape Verde Island (*Cvi-0*): (a) each colour corresponds to different chromosomes in *Arabidopsis thaliana* (nodes (genetic markers) with similar colours belong to the same chromosome; the different edge colours show the positive (red) and negative (blue) partial correlations); (b) 0 pattern of the partial correlation matrix

Table 3. Summary of model fit to the *Arabidopsis thaliana* genotype data

Model	*df*	*Log-likelihood*	*Deviance*	*p-value*
Fitted	237	−1098.75		
Saturated	4005	193.35		
Fitted *versus* saturated	3768		2584.2	1

arrested embryo development resulting in seed abortion, whereas the latter interaction causes root growth impairment.

Furthermore, in addition to these two regions we have discovered a few other trans-chromosomal interactions in the *Arabidopsis thaliana* genome. In particular, two adjacent markers, *c1-13869* and *c1-13926*, in the middle of chromosome 1 interact epistatically with the adjacent markers *c3-18180* and *c3-20729*, at the bottom of chromosome 3. The sign of their conditional correlation score is negative, indicating strong negative epistatic selection during inbreeding. These markers therefore seem evolutionarily favoured to come from different grandparents. This suggests some positive effect of the interbreeding of the two parental lines: it could be that the paternal–maternal combination at these two loci protects against some underlying disorder or that it actively enhances the fitness of the resulting progeny.

5.1.1. Fit of model to Arabidopsis thaliana data

Calculating the deviance statistics D enables us to assess the goodness of fit of the method proposed:

$$D = -2\{l_{\mathrm{m}}(\hat{\Theta}) - l_{\mathrm{s}}(\hat{\Theta})\},$$

where $l_{\mathrm{m}}(\hat{\Theta})$ and $l_{\mathrm{s}}(\hat{\Theta})$ denote the log-likelihood of the observations for the fitted model and the saturated model respectively.

In our modelling framework, the log-likelihood of the fitted model corresponds to the $l_Y(\hat{\Theta}_\lambda)$ that we obtain from equation (11). Taking out the penalty term from equation (4) we obtain the non-penalized log-likelihood of the saturated model, as follows:

$$l_{\mathrm{s}}(\hat{\Theta}) = l_Y(\bar{R}) \cong -\frac{n}{2}\log|\bar{R}| - \frac{1}{2}np,$$

where $\bar{R}$ is the estimated covariance matrix that can be calculated through either Gibbs sampling or approximation approaches in Sections 3.2.1 or 3.2.2. Table 3 shows that the model proposed fits the *Arabidopsis thaliana* data. The χ^2-test with 3768 degrees of freedom gives a high p-value, indicating that there is no reason to suspect a lack of fit.

5.1.2. Evaluating the estimated network in Arabidopsis thaliana

We use a non-parametric bootstrapping approach to determine the uncertainty that is associated with the estimated edges in the precision matrix in *Arabidopsis thaliana*. We generate 200 independent bootstrap samples—as described in Section 3.4—from the genotype data of the *Col-0* and *Cvi-0* cross. For each 200 bootstrap samples we apply the proposed Gaussian copula graphical model as described in Section 3. Furthermore we calculate the frequency of each entry in $\hat{\Theta}^b_{\hat{\lambda}}$ ($b = 1, \ldots, 200$) that have the same sign as the estimated $\hat{\Theta}_{\hat{\lambda}}$ in the original *Cvi-0* and *Col-0* genotype data set. In Fig. 6 we report the corresponding relative frequencies for a sign match of each link across the 200 bootstrap samples.

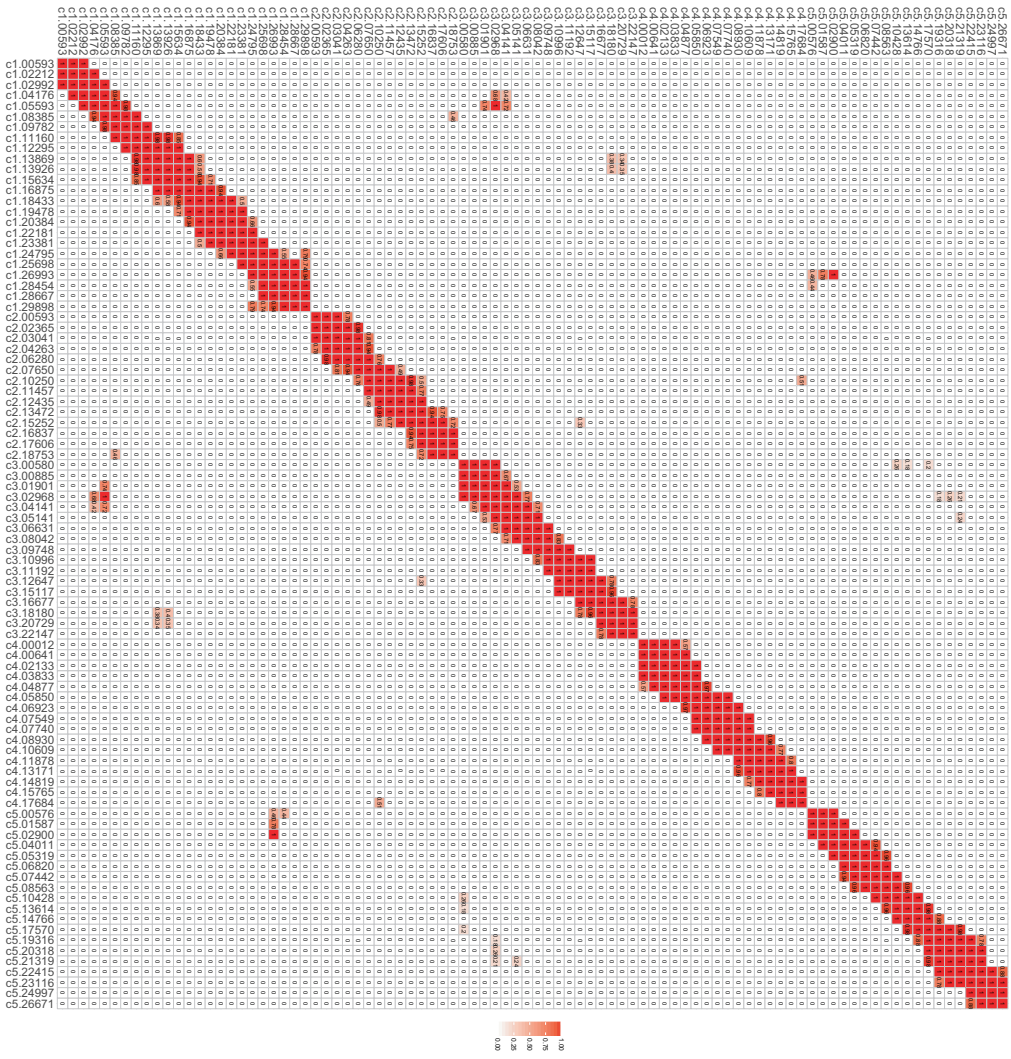


Fig. 6. Uncertainty associated with the estimation of the precision matrix in *Arabidopsis thaliana* using a non-parametric bootstrap: the off-diagonal elements represent the certainty of epistatic interactions between markers in different chromosomes, whereas the 'thick' red diagonal reflects the linkage between neighbouring markers within the chromosomes and shows the linear chromosome structure

Fig. 6 shows the uncertainty that is associated with the epistatic interactions between markers in chromosomes 1 and 5. In particular, in all bootstrap samples we infer a positive epistatic interaction between markers *c1-26993* and *c5-02900*. Also their neighbouring markers interact epistatically. Another region in the *Arabidopsis thaliana* genome that contains epistatic interactions is between chromosomes 1 and 3. In all bootstrap samples, we infer positive epistatic interaction between markers *c1-05593* and *c3-02968*, including their neighbouring markers. Bikard *et al*. (2009) showed that these two regions cause arrested embryo development and root growth impairment in *Arabidopsis thaliana* respectively. In addition to these two confirmed regions we have found other trans-chromosomal regions with potential epistatic interactions. For example, two neighbouring markers in chromosome 1, namely *c1-138669* and *c1-13869*, have quite

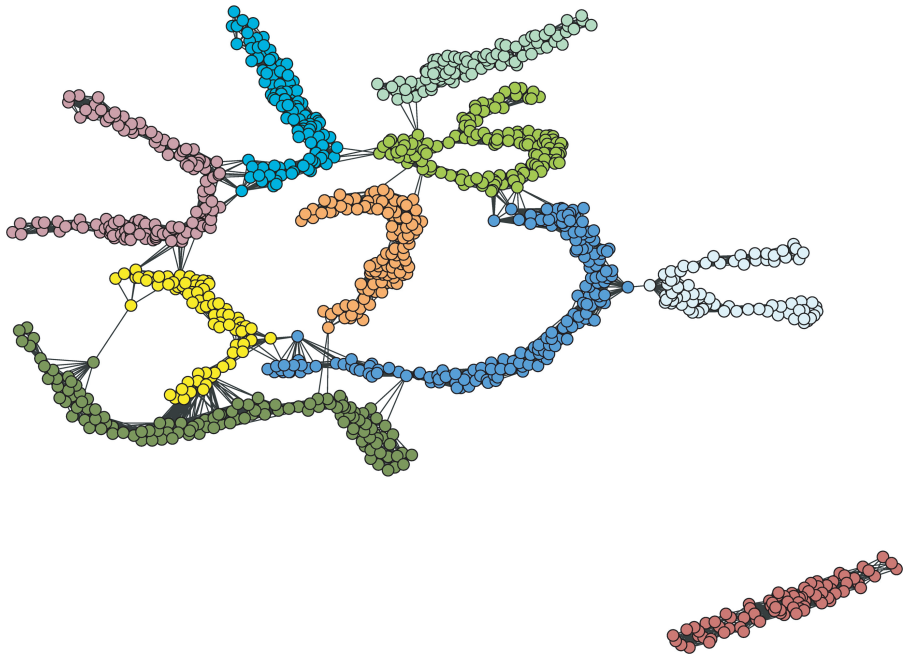


Fig. 7. Inferred network for 1106 markers in the cross between *B73* and *Ki11* in maize by using the approximated method in the Gaussian copula graphical model: , chromosome 1; , chromosome 2; , chromosome 3; , chromosome 4; , chromosome 5; , chromosome 6; , chromosome 7; , chromosome 8; , chromosome 9; , chromosome 10

consistent negative epistatic interactions with the two neighbouring markers in chromosome 3, namely *c3-20729* and *c3-18180*.

5.2. Genetic inbreeding experiment in maize

The nested association mapping initiative in maize populations is designed to reveal the genetic structure of underlying complex traits in maize (McMullen *et al.*, 2009; Rodgers-Melnick *et al.*, 2015). As part of this study, an inbred *Ki11* maize line was crossed with the *B73* reference line. These genotype data contain 1106 markers genotyped for 193 individuals. The $B73 \times Ki11$ RIL is a diploid population with $k = 3$ possible genotypes. We applied our proposed approximation method to the $B73 \times Ki11$ sample, aiming to reveal genetic regions in the maize genome that interact epistatically and may lead to maize disease, e.g. growth impairments, lower fertility or sterility. Exploring genomic signatures of such high dimensional epistatic selection has so far been left unexplored in previous analyses of this essential crop. Fig. 7 shows that certain loci on different chromosomes do not segregate independently of each other. For instance, marker *PZA02117.1* in chromosome 1 interacts with markers *PZA02148.1* in chromosome 6, and marker *PZA00545.26* in chromosome 5 interacts with the three adjacent markers *PZA00466.1*, *PZA01386.3* and *PZA02344.1* in chromosome 9. Existence of such trans-chromosomal conditional dependences indicates marker–marker associations that are possibly due to epistatic selection. Statistically speaking, conditional dependence relationships between physically unlinked pairs of genetic regions contribute to some disorders in this crop that affect its viability.

6. Discussion

Epistatic selection involves the synergistic effects of combinations of genotypes at two or more loci. Epistatic selection can create LD between loci, within and across chromosomes. LD distortions may therefore point to genomic regions undergoing selection. Epistasis is widespread but it may often go undetected because of a lack of statistical power due to testing multiple hypotheses in a possibly very high dimensional setting and due to computational challenges which relate to dealing with missing genotypes and the large solution space that needs to be explored.

In this paper we have introduced an efficient alternative method based on Gaussian copula graphical models that models epistasis as sparse dependences in a high dimensional setting. It is important to remember that this model is the simplest possible multivariate ordinal model as it uses the least number of parameters, $p(p-1)/2$, to describe the full multivariate dependence structure. The method proposed can handle missing genotype values and it captures the conditionally dependent short-range and long-range LD structure of genomes and thus reveals 'aberrant' marker–marker associations that are due to epistatic selection rather than gametic linkage. Polygenic selection on loci that act additively can easily be detected on the basis of strong allele–frequency distortions at individual loci. Epistatic selection, by contrast, does not produce strong locus-specific distortion effects but instead leads to pairwise allele frequency changes.

The method proposed explores the conditional dependences between large numbers of genetic loci in the genome. To obtain a sparse representation of the high dimensional genetic epistatic network, we implement an l_1-penalized likelihood approach. Other extensions of Gaussian graphical models have also been proposed. Vogel and Fried (2011) extended Gaussian graphical models to elliptical graphical models, whereas Finegold and Drton (2009) provided a latent variable interpretation of the generalized partial correlation graph for multivariate t-distributions. They also employed an EM-type algorithm for fitting the model to high-dimensional data.

In the application of our method to an *Arabidopsis thaliana* recombinant inbred line, we discovered two regions that interact epistatically, which had previously been shown to cause arrested embryo development and root growth impairments. In addition, we employed our method to reveal genomic regions in maize that also do not segregate independently and may lead to lower fertility, sterility, complete lethality or other maize diseases. Although *Arabidopsis thaliana* and maize are both diploid species, nothing in our method is limited to diploids. For triploid species, such as seedless watermelons, or even decaploid species, such as certain strawberries, the same method can be used to find epistatic selection by merely adjusting the parameter k (from 3 to respectively 4 and 11).

Acknowledgements

The authors thank Danny Arends for his helpful suggestions with respect to the software implementation of the method. The authors also acknowledge the contribution of European Cooperation in Science and Technology action CA15109.

Appendix A

The following results on the conditional first and second moments of the truncated normal distribution are used in equations (8) and (9). Suppose that a random variable X follows a Gaussian distribution with mean μ_0 and variance σ_0. For any constant t_1 and t_2, $X|t_1 \leqslant X \leqslant t_2$ follows a truncated Gaussian distribution defined on $[t_1, t_2]$. Let $\epsilon_1 = (t_1 - \mu_0)/\sigma_0$ and $\epsilon_2 = (t_2 - \mu_0)/\sigma_0$; then the first and second moments are

$$E(X|t_1 \leqslant X \leqslant t_2) = \mu_0 + \frac{\phi(\epsilon_1) - \phi(\epsilon_2)}{\Phi(\epsilon_2) - \Phi(\epsilon_1)}\sigma_0,$$

$$E(X^2|t_1 \leqslant X \leqslant t_2) = \mu_0^2 + \sigma_0^2 + 2\frac{\phi(\epsilon_1) - \phi(\epsilon_2)}{\Phi(\epsilon_2) - \Phi(\epsilon_1)}\mu_0\sigma + \frac{\epsilon_1\phi(\epsilon_1) - \epsilon_2\phi(\epsilon_2)}{\Phi(\epsilon_2) - \Phi(\epsilon_1)}\sigma_0^2$$

where Φ^{-1} defines the inverse function of the cumulative distribution function of the standard normal distribution.

References

Abegaz, F. and Wit, E. (2015) Copula Gaussian graphical models with penalized ascent Monte Carlo EM algorithm. *Statist. Neerland.*, **69**, 419–441.

Bateson, W. (1909) *Mendel's Principles of Heredity*. Cambridge: Cambridge University Press.

Behrouzi, P. and Wit, E. C. (2017) netgwas: an R package for network-based genome-wide association studies. *Preprint arXiv:1710.01236*. University of Groningen, Groningen.

Bikard, D., Patel, D., Le Metté, C., Giorgi, V., Camilleri, C., Bennett, M. J. and Loudet, O. (2009) Divergent evolution of duplicate genes leads to genetic incompatibilities within *A. thaliana*. *Science*, **323**, 623–626.

Broman, K. W. (2005) The genomes of recombinant inbred lines. *Genetics*, **169**, 1133–1146.

Colomé-Tatché, M. and Johannes, F. (2016) Signatures of Dobzhansky–Muller incompatibilities in the genomes of recombinant inbred lines. *Genetics*, **202**, 825–841.

Dobra, A. and Lenkoski, A. (2011) Copula Gaussian graphical models and their application to modeling functional disability data. *Ann. Appl. Statist.*, **5**, part 2A, 969–993.

Finegold, M. A. and Drton, M. (2009) Robust graphical modeling with *t*-distributions. In *Proc. 25th Conf. Uncertainty in Artificial Intelligence*, pp. 169–176. Corvallis: Association for Uncertainty in Artificial Intelligence Press.

Foygel, R. and Drton, M. (2010) Extended Bayesian information criteria for Gaussian graphical models. In *Advances in Neural Information Processing Systems* (eds J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel and A. Culotta), pp. 604–612. Red Hook: Curran Associates.

Friedman, J., Hastie, T. and Tibshirani, R. (2008) Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, **9**, 432–441.

Geweke, J. (2005) *Contemporary Bayesian Econometrics and Statistics*. New York: Wiley.

Gibson, G. and Mackay, T. F. (2002) Enabling population and quantitative genomics. *Genet. Res.*, **80**, 1–6.

Guo, J., Levina, E., Michailidis, G. and Zhu, J. (2015) Graphical models for ordinal data. *J. Computnl Graph. Statist.*, **24**, 183–204.

Ibrahim, J. G., Zhu, H. and Tang, N. (2008) Model selection criteria for missing-data problems using the EM algorithm. *J. Am. Statist. Ass.*, **103**, 1648–1658.

Lauritzen, S. L. (1996) *Graphical Models*. Oxford: Clarendon.

Lehmann, E. L. and Casella, G. (2006) *Theory of Point Estimation*. New York: Springer Science and Business Media.

Liu, H., Han, F., Yuan, M., Lafferty, J. and Wasserman, L. (2012) High-dimensional semiparametric Gaussian copula graphical models. *Ann. Statist.*, **40**, 2293–2326.

Liu, H., Lafferty, J. and Wasserman, L. (2009) The nonparanormal: semiparametric estimation of high dimensional undirected graphs. *J. Mach. Learn. Res.*, **10**, 2295–2328.

Liu, H., Roeder, K. and Wasserman, L. (2010) Stability approach to regularization selection (STARS) for high dimensional graphical models. In *Advances in Neural Information Processing Systems* (eds J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel and A. Culotta), pp. 1432–1440. Red Hook: Curran Associates.

Mather, K. and Jinks, J. L. (1982) *Biometrical Genetics*, 3rd edn. London: Chapman and Hall.

McMullen, M. D., Kresovich, S., Sanchez Villeda, H., Bradbury, P., Li, H., Sun, Q., Flint-Garcia, S., Thornsberry, J., Acharya, C., Bottoms, C., Brown, P., Browne, C., Eller, M., Guill, K., Harjes, C., Kroon, D., Lepak, N., Mitchell, S. E., Peterson, B., Pressoir, G., Romero, S., Oropeza Rosas, M., Salvo, S., Yates, H., Hanson, M., Jones, E., Smith, S., Glaubitz, J. C., Goodman, M., Ware, D., Holland, J. B. and Buckler, E. S. (2009) Genetic properties of the maize nested association mapping population. *Science*, **325**, 737–740.

Mohammadi, A., Abegaz, F., van den Heuvel, E. and Wit, E. C. (2017) Bayesian modelling of Dupuytren disease by using Gaussian copula graphical models. *Appl. Statist.*, **66**, 629–645.

Nelsen, R. B. (1999) *An Introduction to Copulas*. Berlin: Springer.

Peterson, C. (1987) A mean field theory learning algorithm for neural networks. *Complx Syst.*, **1**, 995–1019.

Rodgers-Melnick, E., Bradbury, P. J., Elshire, R. J., Glaubitz, J. C., Acharya, C. B., Mitchell, S. E., Li, C., Li, Y. and Buckler, E. S. (2015) Recombination in diverse maize is stable, predictable, and associated with genetic load. *Proc. Natn. Acad. Sci. USA*, **112**, 3823–3828.

Rongling, W. and Li, B. (1999) A multiplicative-epistatic model for analyzing interspecific differences in outcrossing species. *Biometrics*, **55**, 355–365.

Simon, M., Loudet, O., Durand, S., Bérard, A., Brunel, D., Sennesal, F. X., Durand-Tardif, M., Pelletier, G. and Camilleri, C. (2008) Quantitative trait loci mapping in five new large recombinant inbred line populations of Arabidopsis thaliana genotyped with consensus single nucleotide polymorphism markers. *Genetics*, **178**, 2253–2264.

Threadgill, D. W., Hunter, K. W. and Williams, R. W. (2002) Genetic dissection of complex and quantitative traits: from fantasy to reality via a community effort. *Mammln Genome*, **13**, no. 4, 175–178.

Törjék, O., Witucka-Wall, H., Meyer, R. C., von Korff, M., Kusterer, B., Rautengarten, C. and Altmann, T. (2006) Segregation distortion in Arabidopsis c24/col-0 and col-0/c24 recombinant inbred line populations is due to reduced fertility caused by epistatic interaction of two loci. *Theoret. Appl. Genet.*, **113**, 1551–1561.

Vogel, D. and Fried, R. (2011) Elliptical graphical modelling. *Biometrika*, **98**, 935–951.

Vujačić, I., Abbruzzo, A. and Wit, E. (2015) A computationally fast alternative to cross-validation in penalized Gaussian graphical models. *J. Statist. Computn Simuln*, **85**, 3628–3640.

Whittaker, J. (2009) *Graphical Models in Applied Multivariate Statistics*. Chichester: Wiley.

Witten, D. M., Friedman, J. H. and Simon, N. (2011) New insights and faster computations for the graphical lasso. *J. Computnl Graph. Statist.*, **20**, 892–900.

Wu, R. and Li, B. (2000) A quantitative genetic model for analyzing species differences in outcrossing species. *Biometrics*, **56**, 1098–1104.

Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

'Detecting epistatic selection with partially observed genotype data using copula graphical models'.